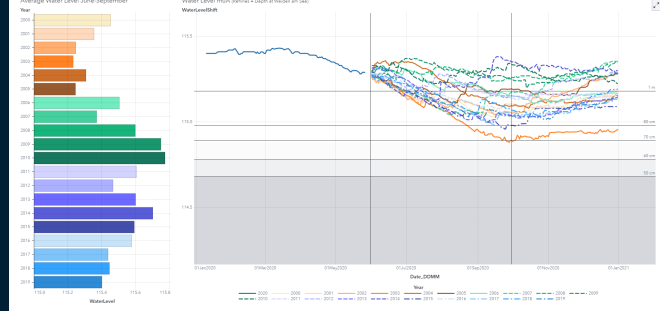




Steht uns das Wasser bis zum Hals? Oder doch nur bis zu den Knien? - Modellierung und Visualisierung der Wasserstandsveränderungen am Neusiedler See mit SAS Viya



Gerhard Svolba, Data Scientist
SAS, DACH Region



Data Scientist @SAS - [Medium](#) [LinkedIn](#) [Github](#) [SAS-Books](#) [SAS Articles](#)
[Youtube](#) [DataPreparation4DataScience](#) [Data Science Use Cases](#)



Überblick

- Den fachlichen Hintergrund verstehen! (Auch in Zeiten von Deep Learning, GPUs, und Container Deployments 😊)
- Feature Engineering (auch bei scheinbar einfachen statistischen Fragestellungen)
- Warum auch ein langjähriger (SAS) Programmierer ab und zu gerne zu visuellen und interaktiven Oberflächen greift
- Manchmal darf es auch eine (einfache) Regressionsanalyse sein
- Bereitstellung und Illustration der Ergebnisse



Simulations-Szenarien für den
Wasserstand des Neusiedler...


SAS Software D-A-CH

[Link](#)

SAS Communities Library

We're smarter together. Learn from this collection of

[Artikel-](#)
[sammlung](#)

Youtube [DataPreparation4DataScience](#) [Data Science Use Cases](#) 

Historischer Tiefststand am Neusiedler See



21. Mai 2020 / BURGENLAND 1619

Neusiedler See trocknet zunehmend aus

Der mittlere Wasserstand liegt unter dem langjährigen Tiefstwert. Nun wird wieder über eine Wasserzufuhr diskutiert



29. Mai 2020 / NEUSIEDLER SEE 333

Grüne und WWF gegen Wasserzufuhr in den Neusiedler See

Landtagsabgeordneter Spitzmüller spricht von einer Gefahr für ein sensibles Natursystem, der WWF gar eine ökologische Katastrophe



26. Mai 2020 / KLIMASTATUSBERICHT 823

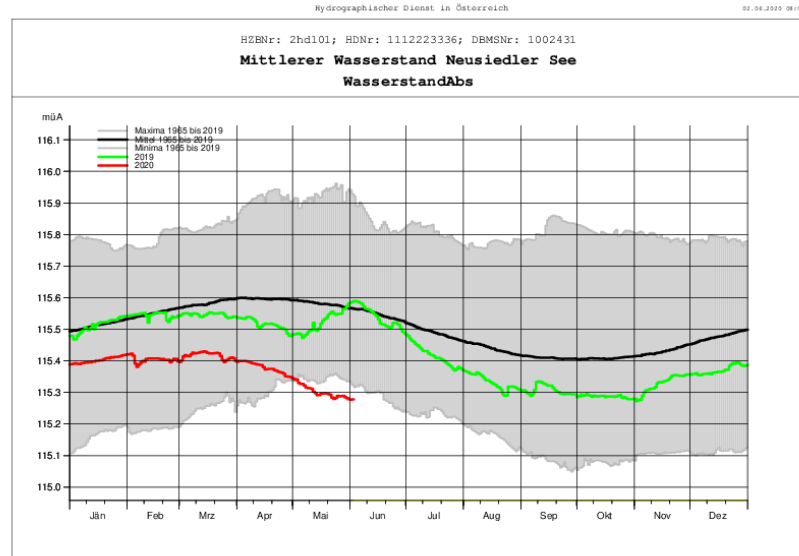
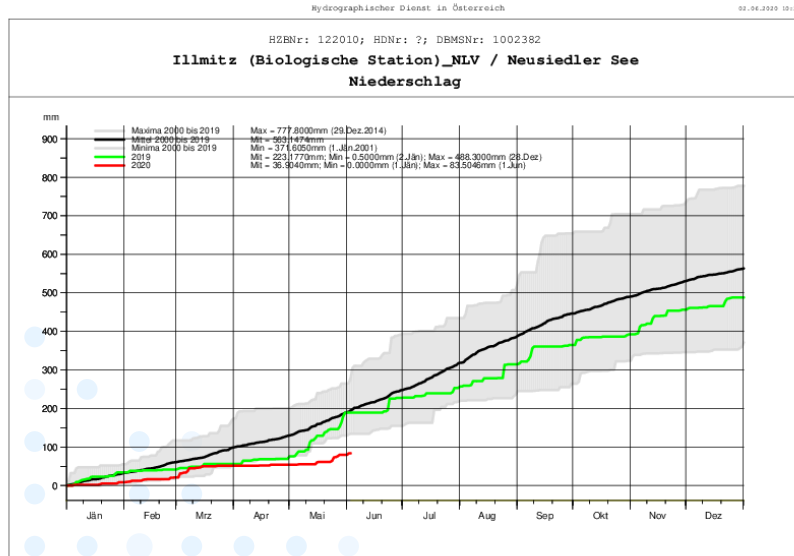
Wie die Klimakrise Österreich erfasst: Dürre, Hitze und Starkregen häufen sich

Hitzesommer und Rekordniederschläge sind längst keine Ausnahme mehr. Künftig werden lange und extreme Wetterperioden laut Klimaforschern zunehmen



Wie kommt das?

- 80% -90% des Wasserhaushalts wird durch Regenwasser gespeist
- Jänner bis Mai 2020 waren extrem trocken
- Karl Maracek (Hydrologie Burgenland): Niederschlag im Winter fördert den Aufbau von Wasserreserven (nur geringe Verdunstung)



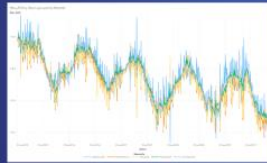
Wie ist der Autor involviert?



Analyse der hydrologischen Daten zum Neusiedler See im Rahmen meiner Lehrveranstaltung an der FH-Burgenland

Task 1.1

Erstellen eines Liniendiagramms pro Station



- Rollenzuordnungen
- Filtern auf Wasserstand (und Messstellen)
- Zuweisung von Farben zu MessStellen
- Override System-Limit for #DataPoints



Task 3.4

Create a cumulative chart per year for the number of hours > 30°C

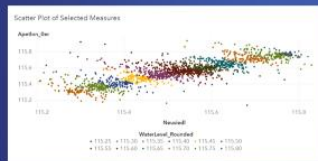


- Similar to task 3.1
- Be careful with the definition GREATER THAN or GREATER EQUAL

SAS

Task 4.3

Detail analysis of the scatter plot



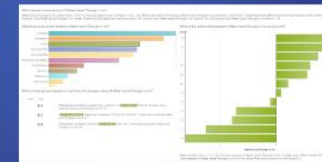
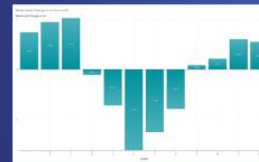
- Subgroup the dots by general (average) water level

SAS

Task 5.2

Study Individual Variables and their Influence

Use Task „Automated Analysis“

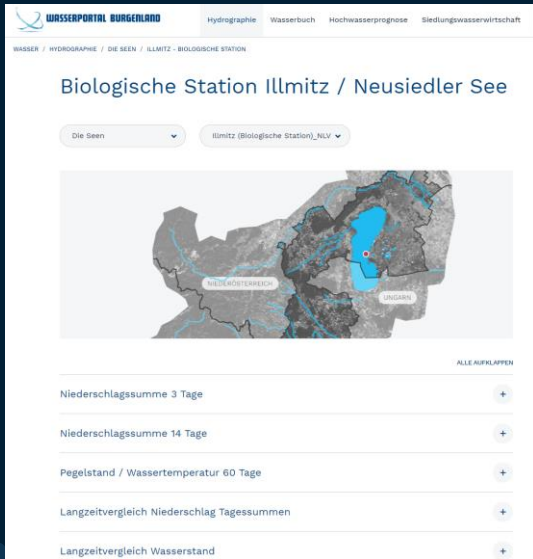


SAS

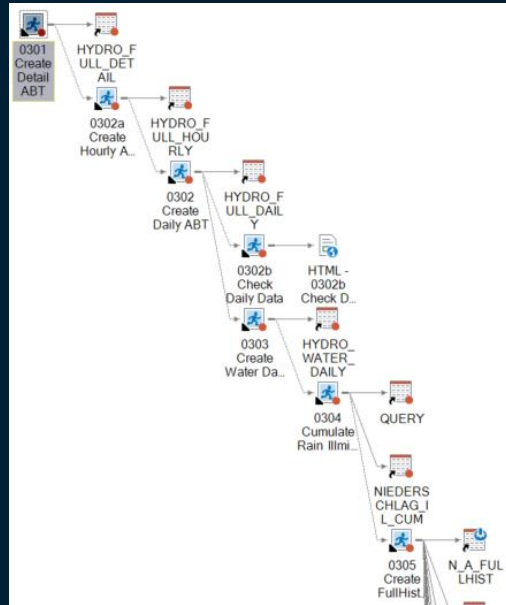
SAS

Überblick über die Vorgangsweise

Laden der Daten von der
Hydrologie Burgenland

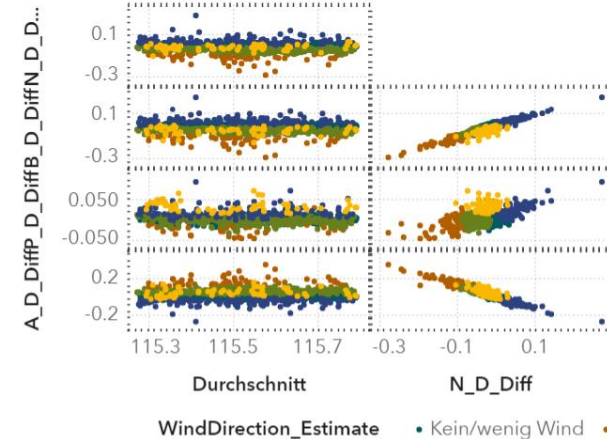


Verknüpfen, Aggregieren
Plausibilitätsprüfung, Aufbereiten



Deskriptive Statistiken, Kausalitäts-
analysen, Simulationen

Scatter Plot of Selected Measures



Wie wollen wir die Daten anordnen?

Wasser - Apetlon

Wasser - Illmitz

Niederschlag - Illmitz

Temperatur - Illmitz

Wasser – Neusiedl (t)

Wasser – Neusiedl
(t+1)

(LONG) Longitudinal, Transactional

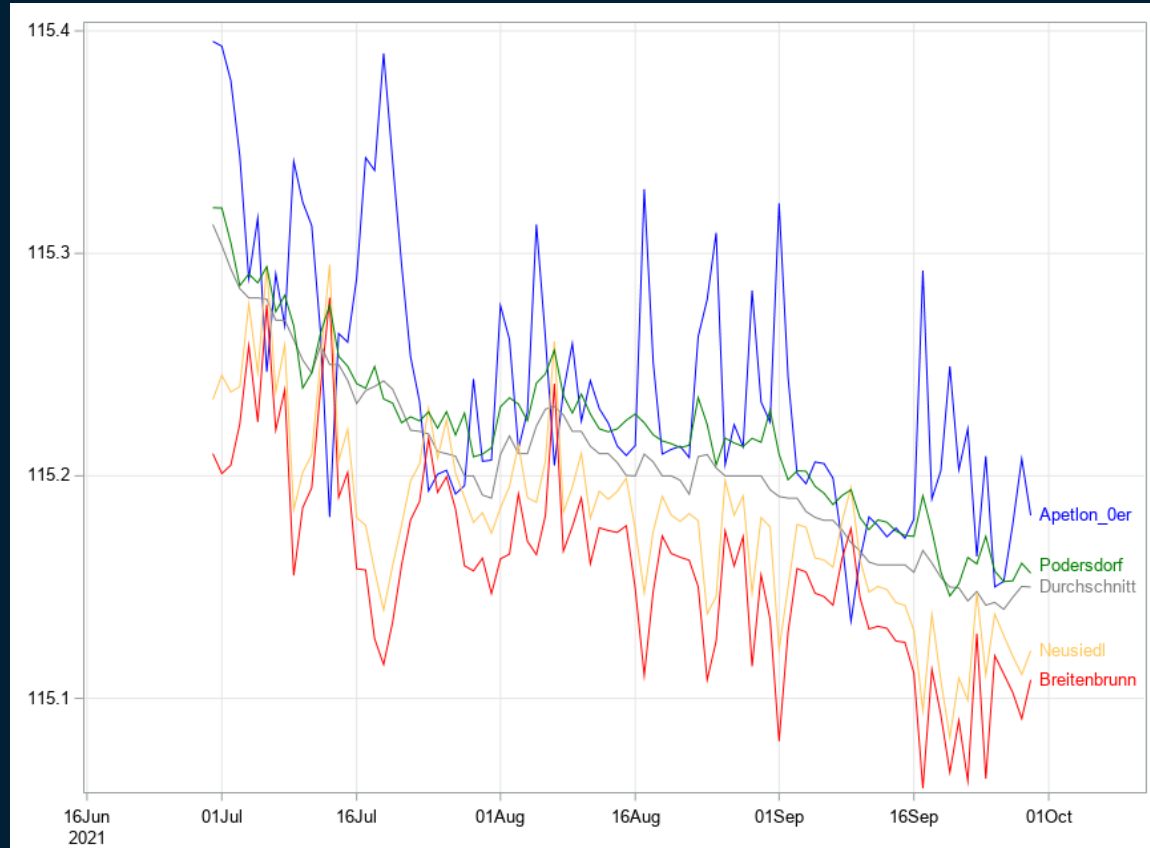
Wasser - Apetlon

Wasser - Illmitz

Wasser - Neusiedl

Temperatur - Illmitz

Niederschlag - Illmitz



Wie sieht unsere Datenstruktur (z.B für Korrelationsanalysen) aus? – WIDE Format

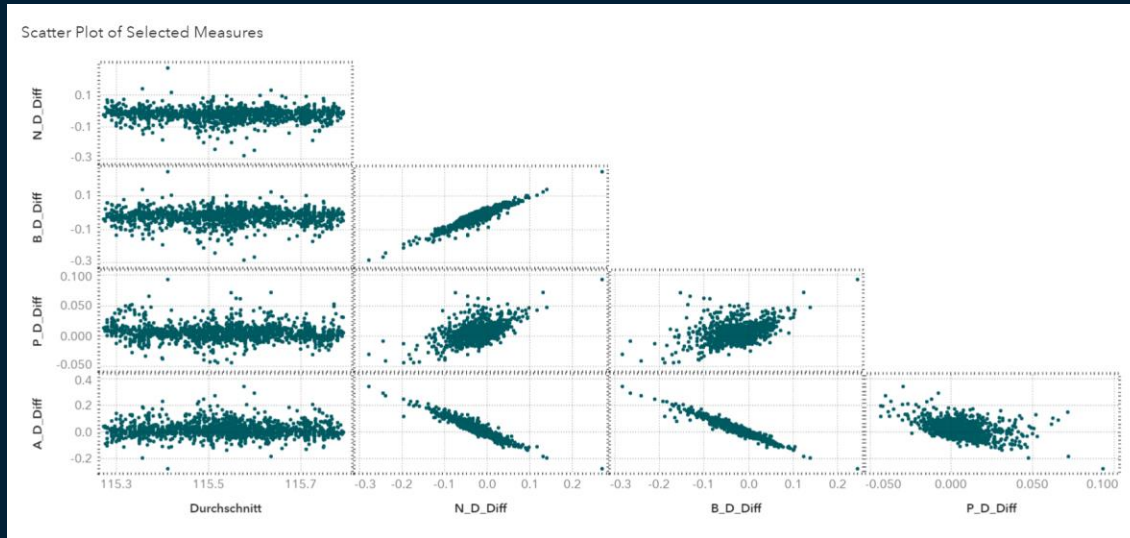
Wasser - Neusiedl

Wasser - Illmitz

Wasser - Apetlon

Temperatur - Illmitz

Niederschlag - Illmitz



Wie sieht unsere Datenstruktur aus?

Mess-Stations Daten	Mess-Größen Daten	Wasser - Neusiedl
Mess-Stations Daten	Mess-Größen Daten	Wasser - Illmitz
Mess-Stations Daten	Mess-Größen Daten	Wasser - Rust
Mess-Stations Daten	Mess-Größen Daten	Temperatur - Illmitz
Mess-Stations Daten	Mess-Größen Daten	Niederschlag - Illmitz

Aggregations-Level ?

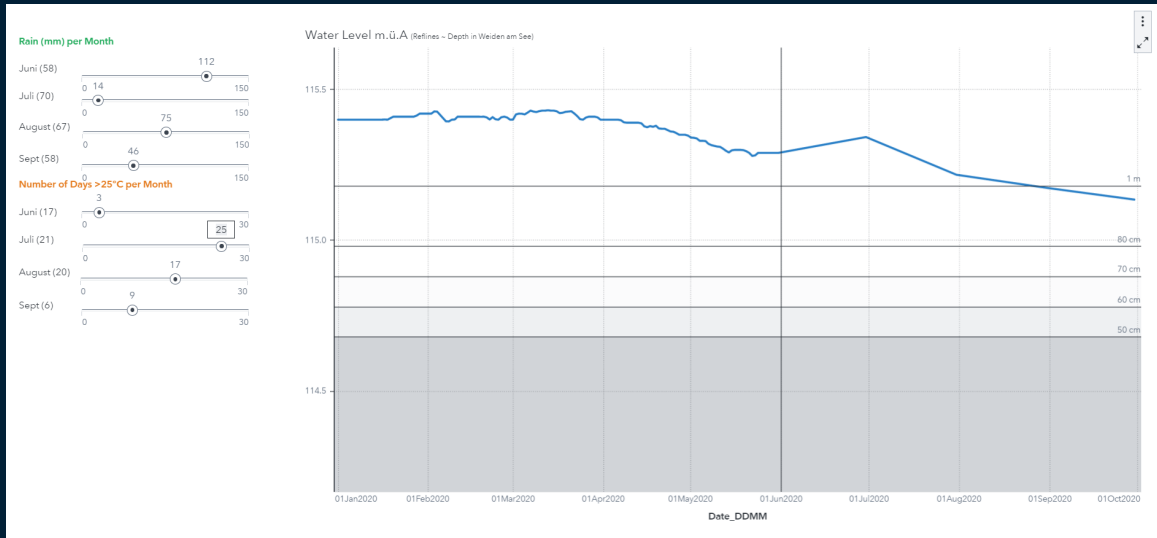
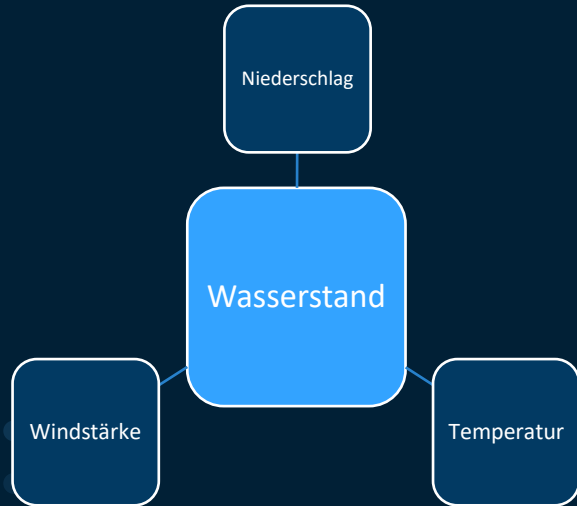
15 min

Stündlich

Täglich

Monat

Use a regression model to explain and quantify the relationship between different factors



Daily data in a WIDE format

Wasser - Neusiedl

Wasser - Illmitz

Wasser - Apetlon

Temperatur - Illmitz

Niederschlag - Illmitz

Feature Engineering

*** Temperature;

Temp_GT30 = (TempMax > 30);

Temp_GT25 = (TempMax > 25);

*** Days Since Last Rain;

retain DaysSinceLastRain 0;

if RainSum > 0 then DaysSinceLastRain = 0;

else DaysSinceLastRain + 1;

*** Water Shift between North/South, East/West;

WShift_N_A = Water_N - Water_A;

WShift_B_P = Water_B - Water_P;

WShift_R_P = Water_R - Water_P;

WShift_B_I = Water_B - Water_I;

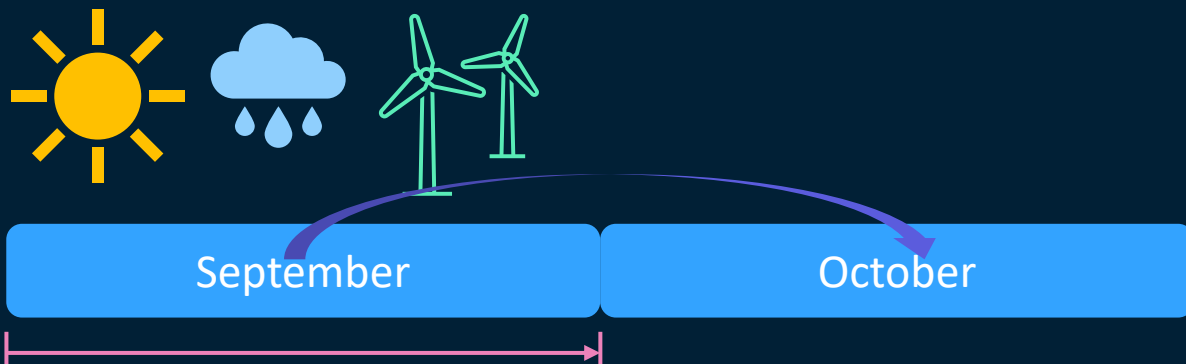
Aggregate from daily to a monthly level

```
select intnx('month',date,0,'Begin') as YYYY format = yymmp7.,  
       month as month,  
       mean(WaterLevel) as WaterLevelMean format=8.2,  
       sum(WaterLevel*(day(date)=1)) as WaterLevel_1ofMonth format =8.2,  
       sum(RainSum) as RainSum format = 8.,  
       sum(RainSum > 0) as Cnt RainDays,  
       max(DaysSinceLastRain) as MaxDaysSinceLastRain,  
       mean(TempMean) as TempMean format = 8.1,  
       sum(Temp_GT30) as Cnt_TmpGT30 ,  
       sum(Temp_GT25) as Cnt_TmpGT25,  
       max(WShift_N_A) as WShift_N_A format=8.2,  
       max(WShift_B_P) as WShift_B_P format=8.2,  
       max(WShift_R_P) as WShift_R_P format=8.2,  
       max(WShift_B_I) as WShift_B_I format=8.2
```

**Observe which
aggregation
statistic you
should use for
each feature!**

Defining the Target Variable: Water_Level_Change

- Sounds simple:
The change between the actual and the next month 😊



Or rather the value on the first day of the respective months?

Comparing the mean of the actual and the next month???

What is the true water level of the lake at a certain point in time?

- Our target variable is slightly biased by the water shift caused by the wind

Wert_AVG by Datum grouped by Messstelle

Wert_AVG by Datum grouped by Messstelle

Wert_AVG

115.30

115.25

115.20

115.15

Datum

Messstelle

Apetlon_0er

Breitenbrunn

Neusiedl

Podersdorf

Durchschnitt

Apetlon_0er

Breitenbrunn

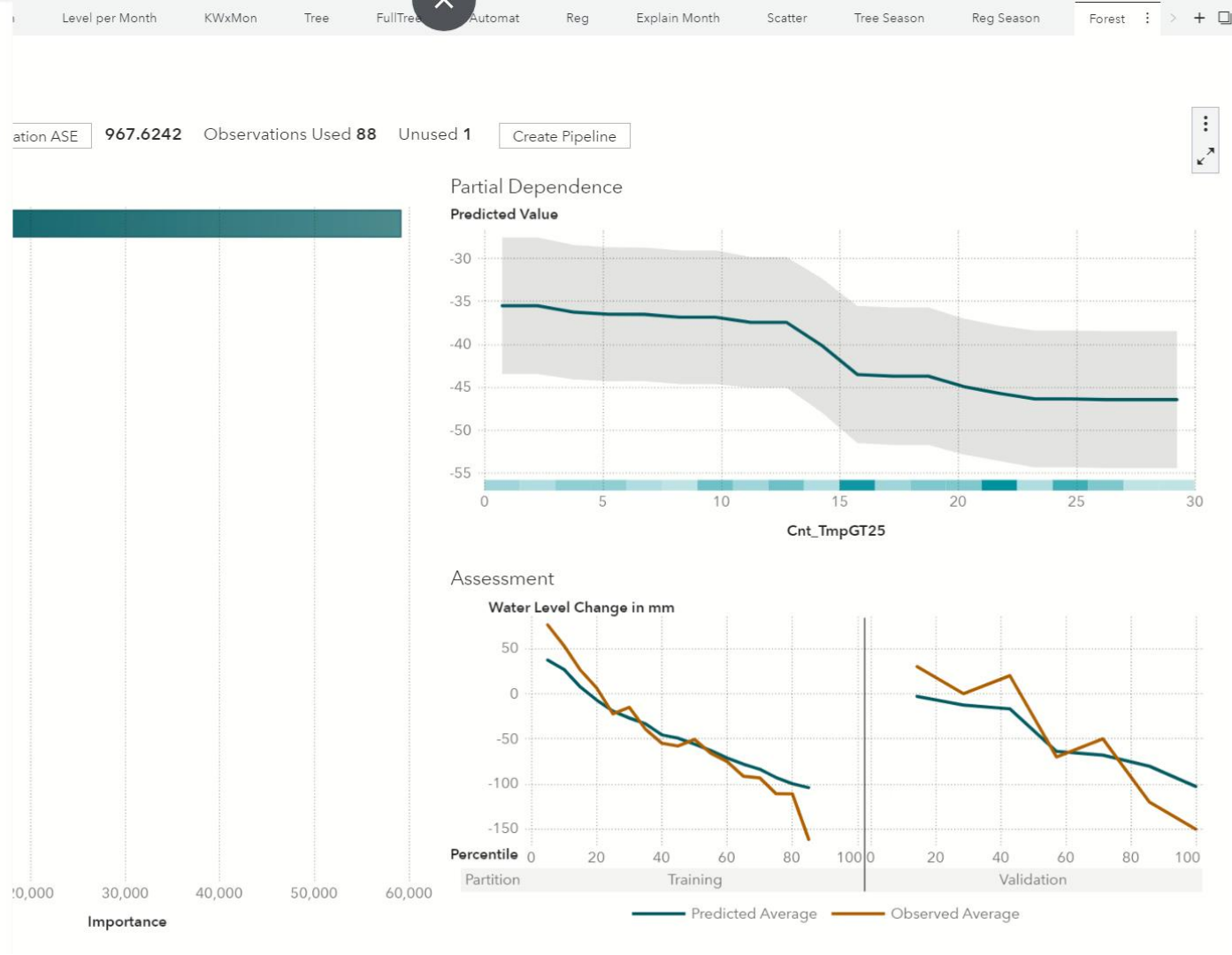
Neusiedl

Podersdorf

Durchschnitt



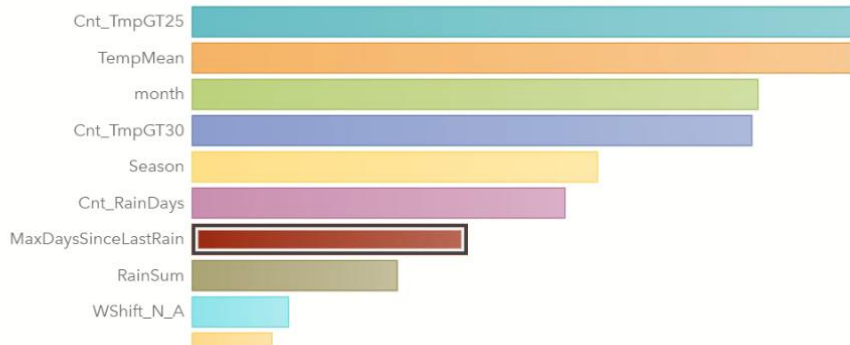
...
Data
HYDRO_ABT_MONTHLY
Filter
+ New data item
Suggest
Outline
Category
Partition_Period - 2
Season - 3
YYMM - 265
Measure
Cnt_RainDays
Cnt_TmpGT25
Cnt_TmpGT30
Frequency
MaxDaysSinceLastRain
month
RainSum
TempMean
Water Level Change in mm
WaterLevel_1ofMonth
WaterLevelMean
WShift_B_I
WShift_B_P
WShift_N_A
WShift_R_P
year
Aggregated Measure
Frequency Percent



What are the characteristics of Water Level Change in mm?

Water Level Change in mm ranges from -161 to 122. Average Water Level Change in mm is -1.5. Most cases (211 of 264) have a Water Level Change in mm between -82 and 70. Cnt_TmpGT25 best differentiates the highest (top 10%) and the lowest (bottom 10%) Water Level Change in mm cases. There is one case that might be an outlier, with Water Level Change in mm below -161.

What factors are most related to Water Level Change in mm?



What are the groups based on MaxDaysSinceLastRain by the average value of Water Level Change in mm?

< High Low >

101

If **MaxDaysSinceLastRain** is less than 5.1, WShift_N_A is between .03 and .08, then the 2 cases have a predicted Water Level Change in mm of 101.

97

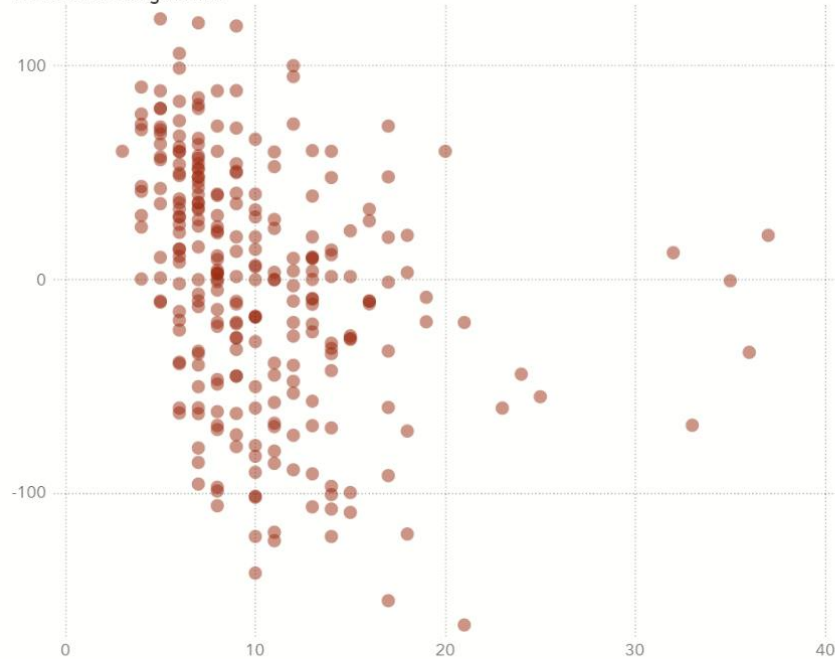
If **MaxDaysSinceLastRain** is less than 6.2, TempMean is between 16 and 19, then the 2 cases have a predicted Water Level Change in mm of 97.

90

If **MaxDaysSinceLastRain** is less than 5.1, RainSum is between 54 and 62, then the 3 cases have a predicted Water Level Change in mm of 90.

What is the relationship between Water Level Change in mm and MaxDaysSinceLastRain?

Water Level Change in mm



Water Level Change in mm may have a strong negative relationship with MaxDaysSinceLastRain. It appears to be a cubic relationship. Average MaxDaysSinceLastRain is 10, and it ranges from 3 to 37.

Use SAS/STAT (or SAS Visual Statistics) to train the model

```
proc glmselect data=monthly_abt_month_sort;  
  model WaterLevelChange =  
    WaterLevelMean RainSum Cnt_Raindays  
    MaxDaysSinceLastRain  
    TempMean Cnt_TmpGT30 Cnt_TmpGT25  
    WShift_N_A WShift_B_P WShift_R_P Shift_B_I  
    /selection=backward;  
where month in (6,7,8,9) and year LE 2019;  
code file="&path.\Hydro_WaterChange_Mod1.0.sas";  
run;
```

Ways how to interact with the SAS Analytic Platform

Graphical User Interface

Visual Interface

Self-service analytical objects with options
Easy integration with Model Studio & Model Manager

Model Studio

Pipelines and Nodes, specific analytical data preprocessing, more options & flexibility, Open Source integration, Easy integration with Model Manager

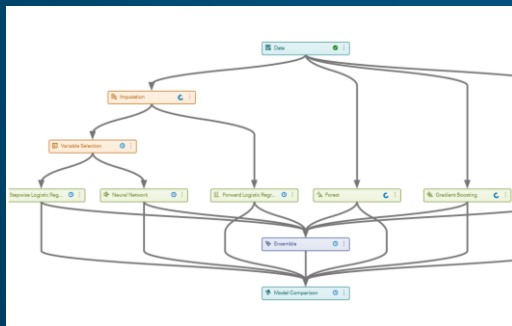
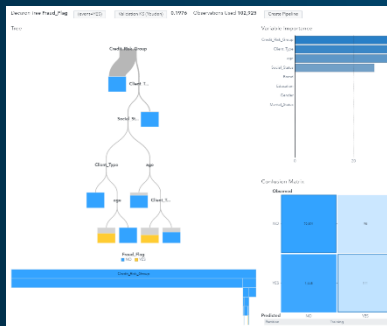
Programming/Coding

SAS

Fully flexible
Procs and actions/actionsets, Open Source integration
Easy integration with Model Manager

OS Language

Fully flexible,
SAS Integration: Access to CAS Procs, actions and actionsets (SWAT)



```
28 proc gradboost data=cas1.fc_review
29   earlystop(tolerance=0 stagnation=5)
30   numline=20 binmethod=BUCKET
31   maxdepth=5
32   maxbranch=2
33   minleafsize=5
34   assignmissing=USEINSEARCH minuseinsearch=1
35   seed=12345
36   printtarget;
37 ;
38 ;
39 partition rolevars=partind_ (TRAIN='1' VALIDATE='0');
40 autotune useparameters=CUSTOM tuningparameters=(
41   lasso(LB=0 UB=10 INIT=0)
42   learningrate(LB=0.01 UB=1 INIT=0.1)
43   nntrees(LB=20 UB=150 INIT=100)
44   ridge(LB=0 UB=10 INIT=0)
45   samplingrate(LB=0.1 UB=1 INIT=0.5)
46   vars_to_try(LB=1 UB=7 INIT=7)
47 )
48 searchmethod=GA objective=KS maxtime=500
49 maxevals=50 maxiters=5 popsize=10
50 targetevents='1'
51 ;
```

```
from swat.reader import reader_html
from swat import *
from pprint import pprint

import matplotlib.pyplot as plt
import pandas as pd
from pandas import *
import numpy as np

import seaborn as sns
get_ipython().magic('matplotlib inline')
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# Provide Connection Information and Upload Data if not yet available
casroot='dch-viya33-emp.viya33.sas.com'
casport=8770
casauth='/j_sauthinfo'
casdata='/j_sauthinfo'
indata_dir='/opt/demodata/sasdata'
indata='new_product_train'
table=indata_dir+"/"+indata+".sas7bdat"
shot_df=pd.read_sas(table)

# Create Some Test Data, SAS Club and Load Actions Sets for Decision Trees
SASClub = CAS(casroot, casport, authinfo='/j_sauthinfo', caslib='casuser')
SASClub.loadactionset(actionset='DecisionTree')

if not SASClub.table.tableExists(table=indata).exists:
    tbl = SASClub.upload_file(indata_dir+"/"+indata+".sas7bdat", casout=('name'+indata))

NOTE: Added action set 'DecisionTree'.
```


Regression Results + Illustration

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	-58.153454	7.587593	-7.66	<.0001
RainSum	1	1.096672	0.062758	17.47	<.0001
Cnt_TmpGT25	1	-3.330063	0.338838	-9.83	<.0001

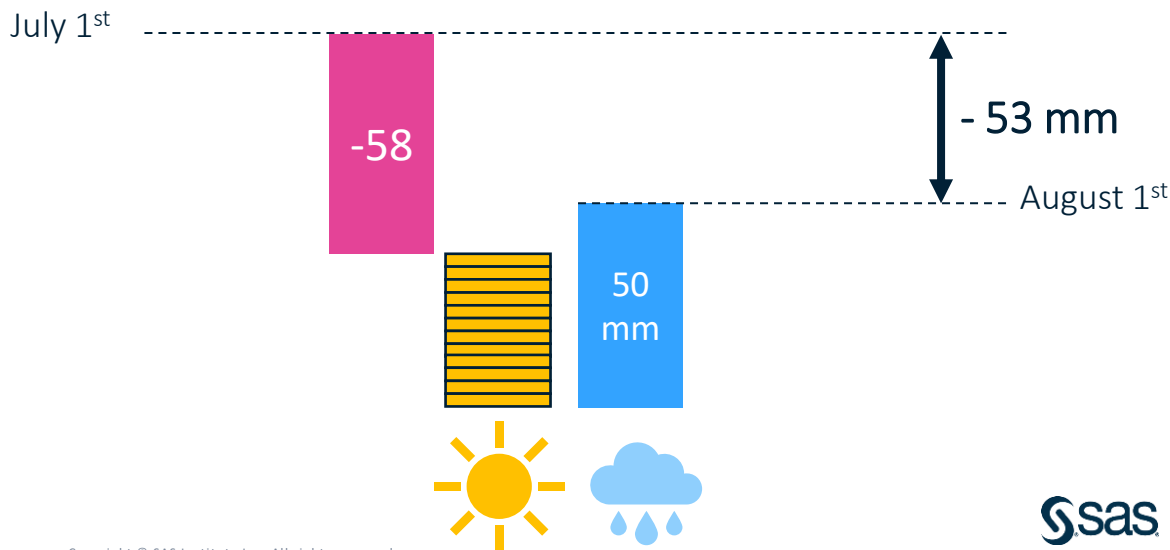
- Average monthly water loss per summer month
- Rain adds to the water level with a factor of ~ 1
- Day $> 25^{\circ}\text{C}$ „costs“ 3.3 mm of water level

Example:

July,

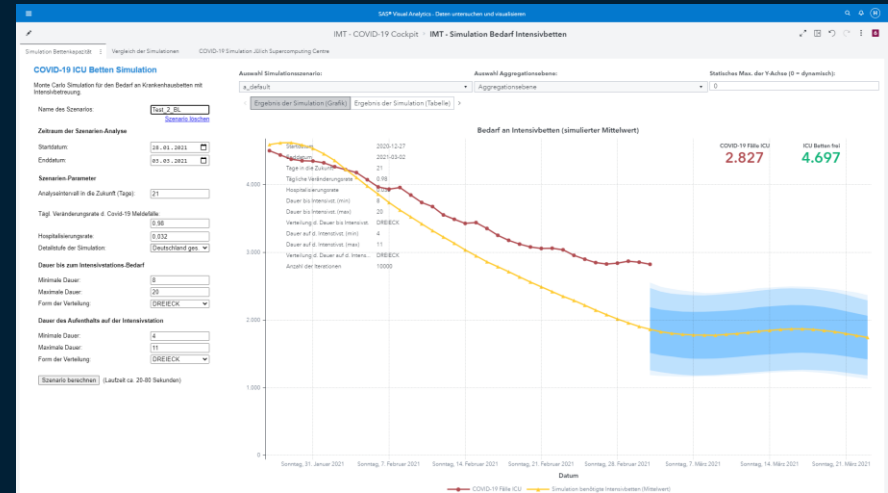
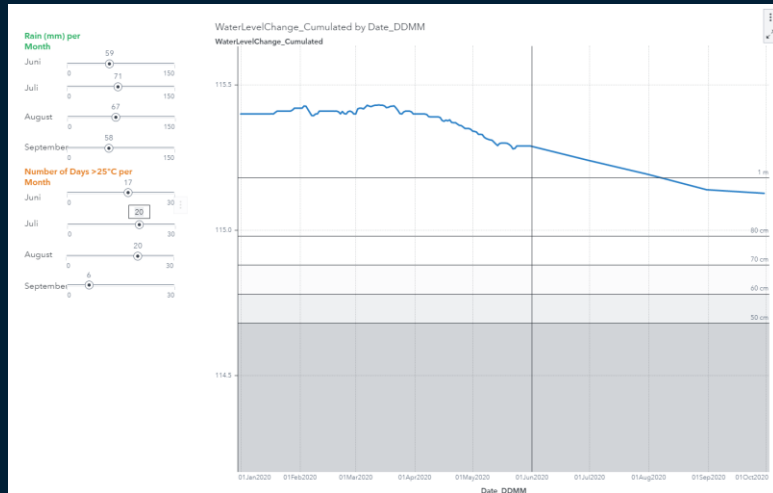
with 12 days $> 25^{\circ}\text{C}$

50 mm rain



Using SAS Visual Analytics to interactively calculate prediction for different scenarios

- Pre-Calculated Scenarios
- Scenario calculated on-the-fly

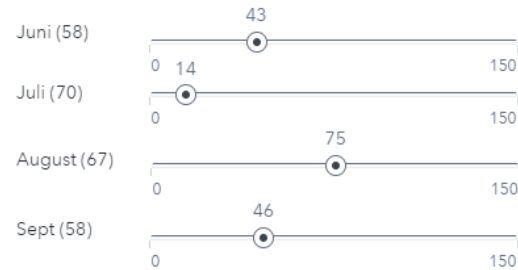


Pre-Calculate outcomes for all possible scenarios and filter according to the selected values

```
Data hydro3.PredWaterLevelChange;  
format month 8. P_WaterLevelChange 8.2;  
do month = 5 to 10;  
  do Cnt_TmpGT25 = 0 to 30;  
    do RainSum = 0 to 200 by 5;  
      %include "\Hydro_WaterChange_Mod1.0.sas";  
      output;  
    end;  
  end;  
end;  
run;
```

month	RainSum	Cnt_TmpGT25	P_WaterLevelChange
6	60	23	-68.94
6	155	14	65.21
6	160	23	40.72
6	125	5	62.28
6	50	9	-33.29

Rain (mm) per Month

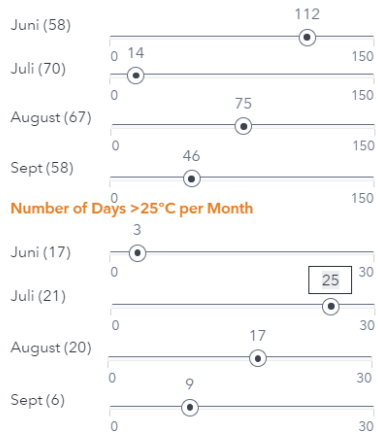


Number of Days >25°C per Month

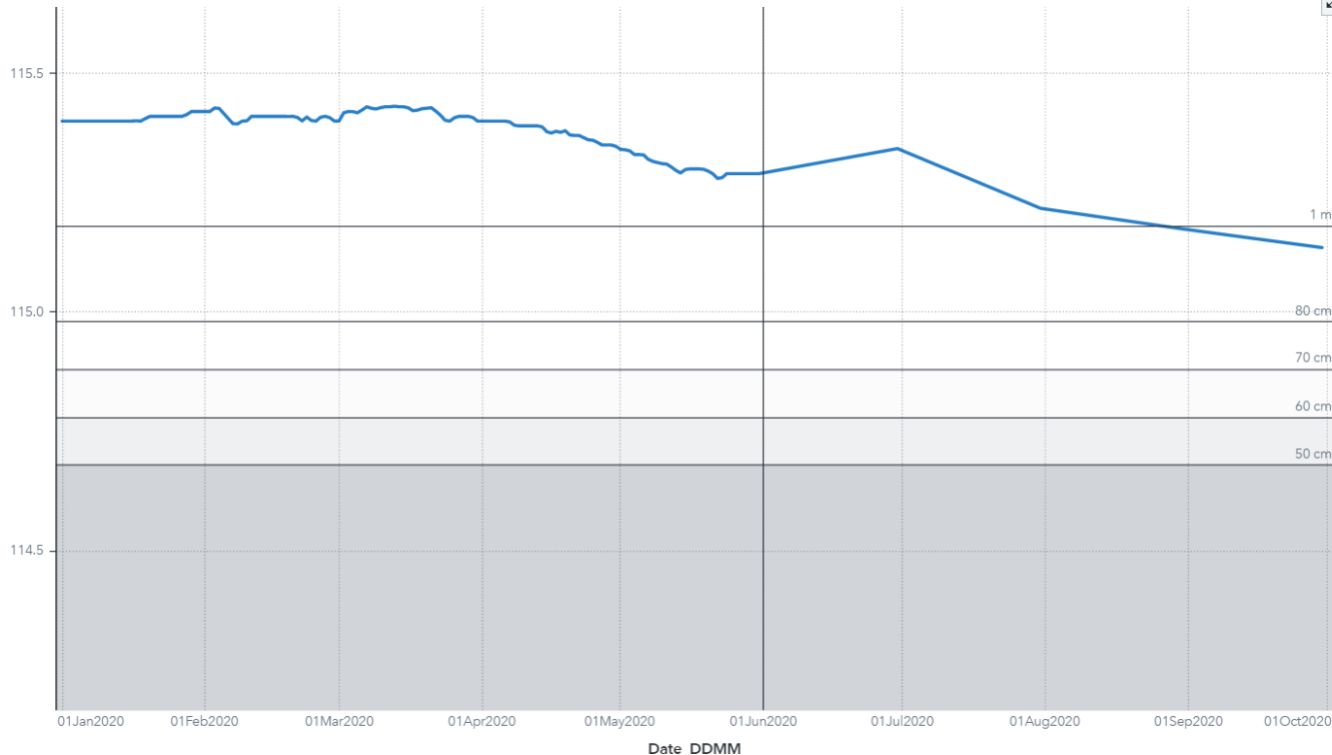


Using SAS Visual Analytics to interactively calculate prediction for different scenarios

Rain (mm) per Month



Water Level m.ü.A (Refines ~ Depth in Weiden am See)



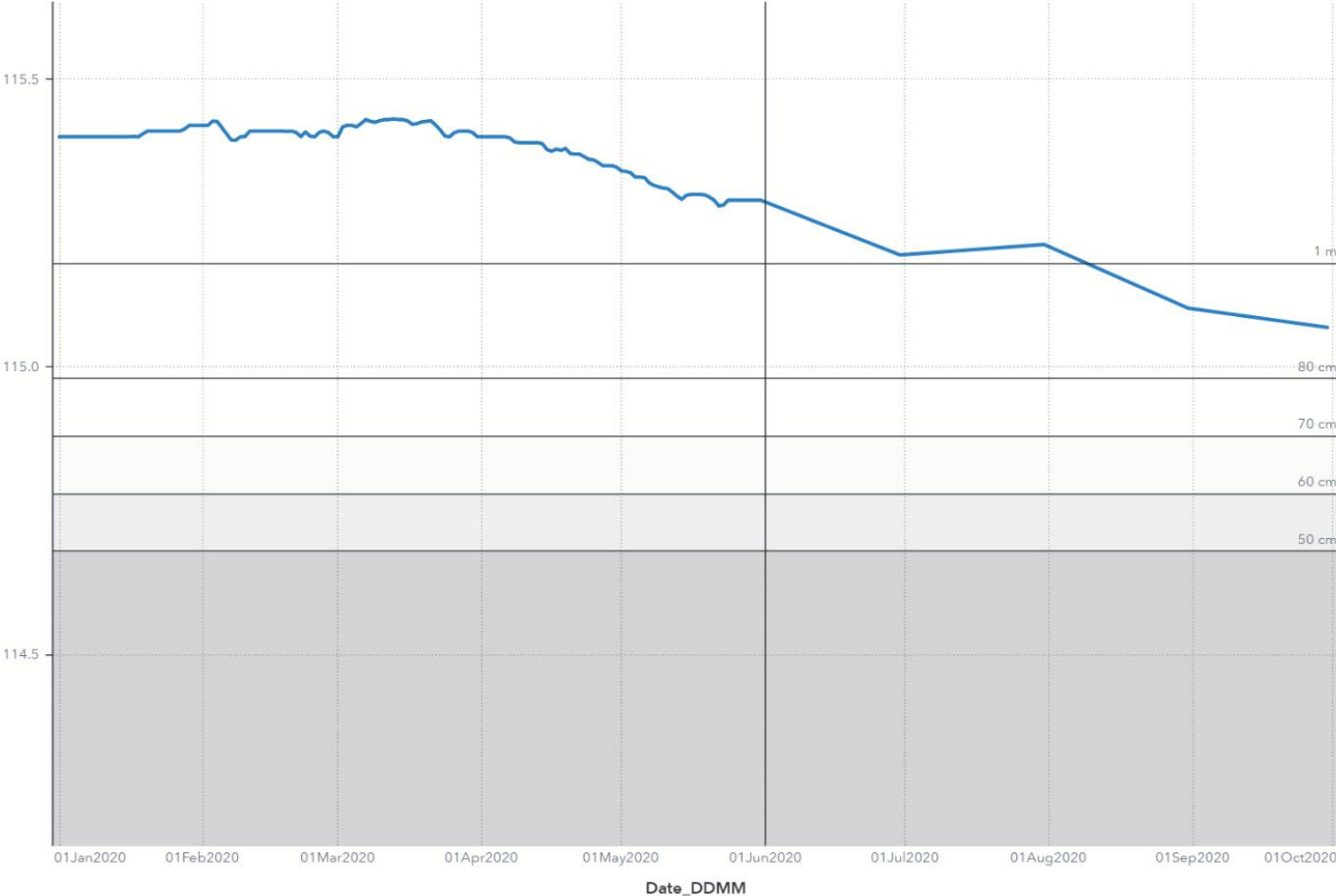
Rain (mm) per Month



Number of Days >25°C per Month



Water Level m.ü.A (Refines ~ Depth in Weiden am See)

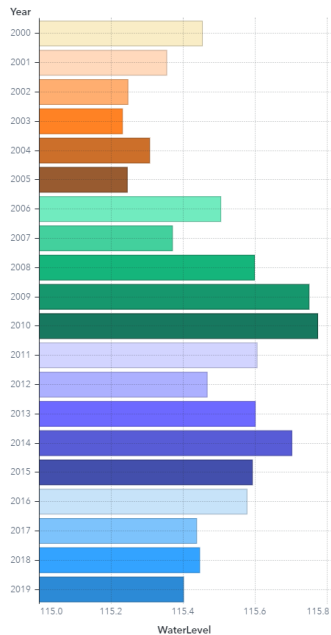


Was wäre wenn,

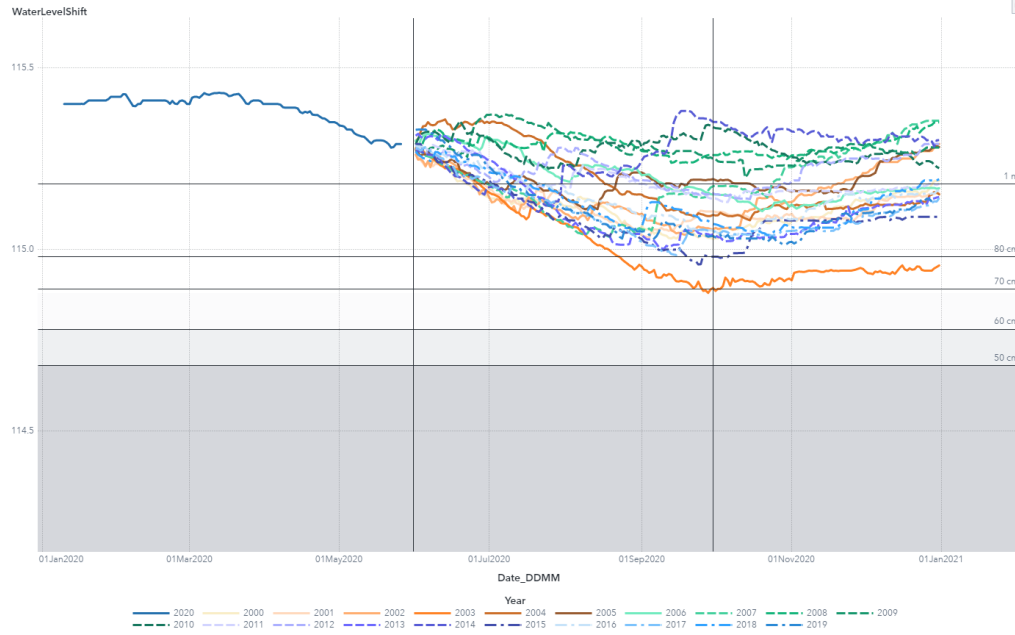
das Wetter im Sommerhalbjahr so ist, wie ...

- im extrem trockenen Jahr 2003,
- im niederschlagsreichen Jahr 2014,
- ...

Average Water Level June-September

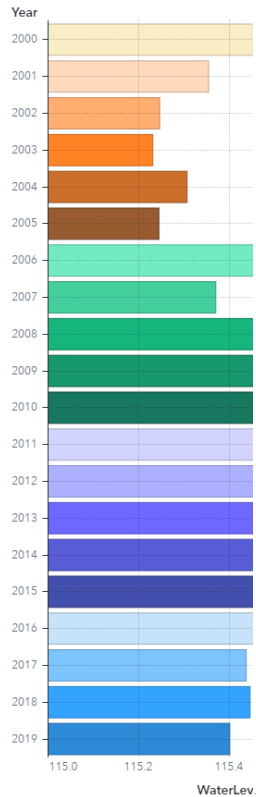


Water Level müA (Refines = Depth at Weiden am See)



Interactive Display in SAS Visual Analytics: Selecting a bar on the left filters line chart on the right

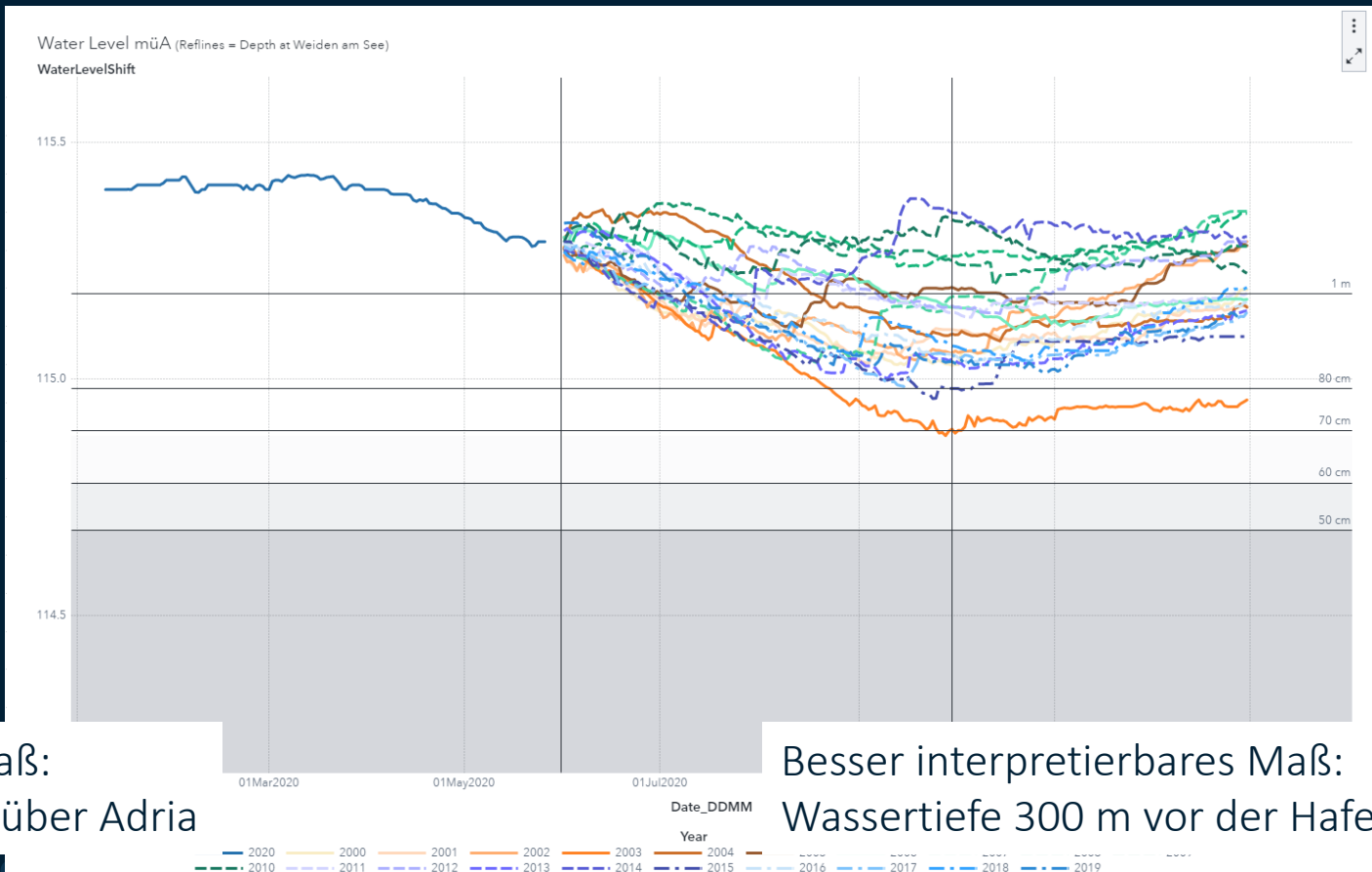
Average Water Level June-Sept



Water Level m.ü.A (Refines ~ Depth in Weiden am See)



Display Tipp: Use interpretable scales at your graph axes



Offizielles Maß:
müA (Meter über Adria)

Besser interpretierbares Maß:
Wassertiefe 300 m vor der Hafenausfahrt

Überblick

- Den fachlichen Hintergrund verstehen! (Auch in Zeiten von Deep Learning, GPUs, und Container Deployments 😊)
- Feature Engineering (auch bei scheinbar einfachen statistischen Fragestellungen)
- Warum auch ein langjähriger (SAS) Programmierer ab und zu gerne zu visuellen und interaktiven Oberflächen greift
- Manchmal darf es auch eine (einfache) Regressionsanalyse sein
- Bereitstellung und Illustration der Ergebnisse



Simulations-Szenarien für den
Wasserstand des Neusiedler...


SAS Software D-A-CH

[Link](#)

SAS Communities Library

We're smarter together. Learn from this collection of

[Artikel-](#)
[sammlung](#)

Youtube [DataPreparation4DataScience](#) [Data Science Use Cases](#) 

Zusammenfassung

- Den fachlichen Hintergrund verstehen!
Für Feature Engineering, Modellbildung und Interpretation.
- Wer ist mein Auditorium? Wissenschaftler, Data Scientists, Segler, ...
- Was ist die optimale Darstellungsart der Ergebnisse für diese Zielgruppe?
- SAS Visual Analytics unterstützt mich (> 25 Jahre SAS) in der Exploration der Daten und der Aufbereitung der Ergebnisse
- Und viele Fachanwender bei der kompletten Datenanalyse

Data Preparation for Data Science

Data
Assembly

Data Quality
for Analytics

Feature
Generation

Gerhard Svolba,
Data Scientist @SAS
<mailto:sastools.by.gerhard@gmx.net>

Articles
and Blogs



Webinars



Tipps &
Tricks



Macros &
Downloads

